The Cook-Levin Theorem

Vassos Hadzilacos

Theorem 8.7 (Cook ’71, Levin ’73) The satisfiability problem for propositional formulas, SAT, is NP-complete.

Proof. It is clear that SAT is in NP (the certificate is a truth assignment, which is short, and the verifier checks that the truth assignment satisfies the formula, which can be done in polynomial time). The more interesting part is the proof that SAT in NP-hard. Take any decision problem $A \subseteq \Sigma^*$ in NP, and let $M_A = (Q, \Sigma, \Gamma, \delta, q_0, h_A, h_R)$ be a nondeterministic Turing machine that decides $A$ in polynomial time, say $p(n)$, where $n$ is the length of the input $x$. Given any $x \in \Sigma^*$ we show how to construct a propositional formula $F_x$ that is satisfiable if and only if $x \in A$; that is, if and only if $M_A$ on input $x$ has an accepting computation $C_0 \vdash C_1 \vdots \vdash C_t$, where $t \leq p(|x|)$. The length of $F_x$ will be polynomial in $|x|$, and so it will be obvious that it can be constructed from $x$ in polynomial time in $|x|$. Fix any $x \in \Sigma^*$ and let $n = |x|$.

The propositional variables involved in $F_x$ describe the state of affairs in the computation of $M_A$ on input $x$ at each “time” $t$ (i.e., after $t$ steps), $0 \leq t \leq p(n)$. (If $t < p(n)$ we will imagine that $M_A$ keeps going until time $p(n)$ without changing its state, head position, or tape contents.) The variables are listed below, along with their intended meaning.

- $S^q_t$, for $0 \leq t \leq p(n)$ and $q \in Q$: At time $t$, $M_A$ is in state $q$.
- $H^i_t$, for $0 \leq t \leq p(n)$ and $1 \leq i \leq p(n) + 1$: At time $t$, the head of $M_A$ is on cell $i$. Note that in $p(n)$ steps the rightmost cell that $M_A$ can reach is $p(n) + 1$.
- $T^a_t$, for $0 \leq t \leq p(n)$, $1 \leq i \leq p(n)$, and $a \in \Gamma$: At time $t$, the cell $i$ of $M_A$’s tape contains symbol $a$. (Note that all cells to the right of cell $p(n)$ can only contain blanks.)

Thus our formula will have $O(p(n)) + O(p^2(n)) + O(p^2(n)) = O(p^2(n))$ variables, i.e., a polynomial number of them. Note that the number of states $|Q|$ and the number of symbols $|\Gamma|$ are constants: these depend on $M_A$, not on the input $x$.

The formula $F_x$ is the conjunction of four subformulas, each expressing a requirement for the computation of $M_A$ on input $x$:

1. Intuitively the subformula $F^1_x$ states that, at each time $t$, the variables describe a coherent state of affairs: $M_A$ is in at most one state, its head is in at most one place, and each tape cell contains at most one symbol. (It will follow from this and other subformulas that $M_A$ is actually in exactly one state, the head in exactly one place, and each cell has exactly one symbol.) This is expressed as follows:

$$F^1_x = \bigwedge_{0 \leq t \leq p(n)} \left(\bigwedge_{p \neq q \in Q} \left(\neg S^q_t \lor \neg S^p_t\right)\right) \land \left(\bigwedge_{1 \leq i \leq p(n)} \left(\neg H^i_t \lor \neg H^j_t\right)\right) \land \left(\bigwedge_{1 \leq i \leq p(n)} \bigwedge_{a \neq b \in \Gamma} \left(\neg T^a_t \lor \neg T^b_t\right)\right).$$

2. Intuitively the subformula $F^2_x$ states that the computation of $M_A$ on $x$ starts well: At time 0, $M_A$ is in its initial state $q_0$, its head is on cell 1, and the tape contains the input $x$ in the first $n$ cells and blanks in cells $n + 1..p(n)$. This is expressed as follows: Let $x = a_1 a_2 \ldots a_n$, where each $a_i \in \Sigma$.

$$F^2_x = S_0^{q_0} \land \left(\bigwedge_{1 \leq i \leq n} T^a_0\right) \land \left(\bigwedge_{n+1 \leq i \leq p(n)} T^b_0\right).$$
(3) Intuitively the subformula $F^3_x$ states that the computation of $M_A$ on $x$ ends well: At time $p(n)$, $M_A$ is in its accept state $h_A$. This is expressed as follows:

$$F^3_x = S^h_{p(n)}.$$  

(4) Finally, the subformula $F^4_x$ is the conjunction of two other formulas, $F^{4a}_x$ and $F^{4b}_x$: $F^{4a}_x$ states that as long as the current state of the TM is not the accept state, in each step the state of affairs changes according to the (nondeterministic) transition function of $M_A$ (which, recall, is nondeterministic), until $M_A$ accepts. $F^{4b}_x$ states that after reaching the accept state, things don’t change. These are expressed as follows:

$$F^{4a}_x = \bigwedge_{0 \leq t \leq p(n)} \bigwedge_{q \in Q-\{h_A\}} \bigwedge_{1 \leq i \leq p(n)} \bigwedge_{a \in \Gamma} \left( \neg S^q_t \vee \neg H^i_t \vee \neg T^{ia}_t \right) \vee \left( \bigvee_{(p,b,d) \in \delta(q,a)} (S^p_{t+1} \wedge H^i_{t+1} \wedge T^{ib}_{t+1}) \right).$$

where

$$d = \begin{cases} 1, & \text{if } D = R \\ -1, & \text{if } D = L \text{ and } i \neq 1 \\ 0, & \text{otherwise} \end{cases}$$

$$F^{4b}_x = \bigwedge_{0 \leq t \leq p(n)} \bigwedge_{1 \leq i \leq p(n)} \bigwedge_{a \in \Gamma} \left( \neg S^h_{t+1} \vee \neg H^i_{t+1} \vee \neg T^{ia}_{t+1} \right) \vee \left( (S^h_{t+1} \wedge H^i_{t+1} \wedge T^{ia}_{t+1}) \right).$$

Note that if the machine reaches the reject state, the formula $F^{4a}_x$ becomes unsatisfiable because there is no next state.

So, the overall formula $F_x$ is

$$F_x = F^1_x \wedge F^2_x \wedge F^3_x \wedge (F^{4a}_x \wedge F^{4b}_x).$$

Given the semantics of each subformula, $F_x$ asserts that $M_A$ on input $x$ has an accepting computation. In other words, $F_x$ is satisfiable if and only if $x \in A$.

If one were to prove this fact in more detail (which we will not), one would argue as follows: For the “only if” direction, from an accepting computation

$$C_0 \models C_1 \models \ldots \models C_\ell$$

of $M_A$ on $x$ we would define truth values for all the variables based on their intended meaning (e.g., $S^q_t$ would be true if $C_t$ contains the state $q$ and false otherwise). We would then argue that this truth assignment satisfies all subformulas and therefore their conjunction $F_x$. (In the event $\ell < p(n)$ we would also define the truth values of the variables corresponding to times $t, \ell < t \leq p(n)$, to be equal to their values at time $\ell$.) For the “if” direction, starting from a satisfying truth assignment to the variables of $F_x$ we would define configurations that correspond to an accepting computation of $M_A$ on $x$, showing that $x \in A$.

Now let us calculate the length of $F_x$, measured as the number of variable occurrences. Recalling that $|Q|$ and $|\Gamma|$ are constants we see that the lengths of $F^1_x$, $F^2_x$, $F^3_x$, $F^{4a}_x$, and $F^{4b}_x$ are, respectively $O(p^3(n))$, $O(p(n))$, $O(1)$, $O(p^2(n))$, and $O(p^2(n))$. (For $F^{4a}_x$ note that the maximum number of choices due to the nondeterminism of $M_A$ is $(|Q|-2) \cdot |\Gamma|$, which is constant.) Therefore $F_x$ is of polynomial size in $n$ (the length of the input $x$), and obviously can be constructed from $x$ in polynomial time.

So there is a polynomial time mapping reduction from any decision problem in NP to SAT.  

The formula $F_x$ in the proof of the Cook-Levin theorem is almost in conjunctive normal form (CNF). Only the subformula $F^{4a}_x$ is not. We will now show that we can put that subformula in CNF without
sacrificing the polynomial size of the resulting formula. To simplify the notation, $F^{3a}_x$ is a conjunction of formulas of the form

$$\phi = (\ell_1 \lor \ell_2 \lor \ell_3) \lor (\bigvee_{i=1}^{k} (\ell_{i1} \land \ell_{i2} \land \ell_{i3}))$$

where $\ell_1, \ell_2, \ell_3$ and the $\ell_{ij}$s are literals, and $k$ is the number of choices of $M_A$’s nondeterministic transition function, a constant. We can put $\phi'$ in CNF by applying the distributive law of conjunctions over disjunctions, resulting in the following equivalent formula:

$$\phi'' = \bigwedge_{\pi \in \{1, 2, 3\}^{\{1, 2, \ldots, k\}}} (\ell_{1\pi(1)} \lor \ell_{2\pi(2)} \lor \ldots \lor \ell_{k\pi(k)}).$$

(If $X$ and $Y$ are sets, $Y^X$ denotes the set of all functions from $X$ to $Y$. Thus a function $\pi \in \{1, 2, 3\}^{\{1, 2, \ldots, k\}}$ maps each $i = 1, 2, \ldots, k$ to 1, 2, or 3. Intuitively, $\pi$ selects one of the three literals of each clause in $\phi'$.) By replacing $\phi'$ by $\phi''$ in $\phi$ and applying the distributive law once more (now of disjunctions over conjunctions) we get that $\phi$ is equivalent to the following formula

$$\psi = \bigwedge_{\pi \in \{1, 2, 3\}^{\{1, 2, \ldots, k\}}} (\ell_1 \lor \ell_2 \lor \ell_3 \lor \ell_{1\pi(1)} \lor \ell_{2\pi(2)} \lor \ldots \lor \ell_{k\pi(k)}).$$

which is in CNF.

The length of $\phi$, measured as the number of variable occurrences, is $3k + 3$, whereas that of $\psi$ is $3^k(k + 3)$. But recall that $k$ is a constant and therefore the size of $F_x$, with $F^{3b}_x$ replaced by an equivalent CNF formula as above, remains polynomial in the size of the input $x$. Therefore we have:

**Corollary 8.8** The satisfiability problem for CNF formulas, CNFSAT, is $NP$-complete.